Can corpus-derived collocation metrics reflect expectations in language processing?

Kyla McConnell & Alice Blumenthal-Dramé

Comprehenders track distributional statistical information online and use this to inform expectations (Kuperberg & Jaeger, 2016). This can be seen in the processing of collocations, two or more words that appear together more often than would be expected by chance. Consider the case of the modifier-noun collocation *vast majority*: where *vast* appears, there is a strong likelihood that *majority* will follow.

In the corpus linguistic and psycholinguistic communities, different metrics have been put forward to quantify association strengths between collocates. Certain metrics have predominantly served as tools in theoretical linguistics, lexicography, and applied linguistics, whereas others have mainly been used to model probabilistic language processing in information-theoretic terms (Evert, 2009; Hale, 2016). But are these metrics psychologically realistic? That is, can they reflect expectations in online language processing?

This self-paced reading study assessed the reading times for the second word in collocated modifier-noun bigrams like *vast majority*. Six of the most common corpus linguistic association scores – MI, MI3, Dice coefficient, T-score, Z-score and log-likelihood – were pitted against predictors of lexical processing cost that are widely established in the psycholinguistic and cognitive linguistic communities, respectively: log-transformed forward/backward transition probability (logForwardTP/logBackwardTP) and bigram frequency (logBigramFreq). In comparing traditional corpus-based metrics with metrics derived from transition probabilities, we bring corpus linguistics together with information-theoretic models. By adding log bigram frequency to the comparison, we contrast these two approaches with recent work in which multi-word units, or chunks, have been suggested to constitute the primitive building blocks of language (i.e. Arnon & Snider, 2010).

123 native speakers of English, recruited online, read 91 critical sentences and 157 filler sentences. Critical sentences contained one modifier-noun bigram embedded in a neutral sentence head and followed by a three-word spillover region (i.e., *Connor was informed about the* <u>*great deal*</u> *on designer jeans*). Association scores for each bigram were extracted from the British National Corpus (BNCweb QCP-Edition). Reading times to the noun were analyzed with mixed-effects models with one association score as a fixed effect per model. Models were compared for goodness of fit using the AIC (Akaike information criterion). Raw RTs under 100ms and above 2000ms, as well as logRTs outside of 3 SD from each participant's mean were excluded as outliers (1.85% data loss) and critical word (noun) frequency was controlled for across models.

Results showed that none of the six traditional corpus linguistic metrics patterned significantly with log-transformed reading times to the noun at the Bonferroni-corrected significance level of 0.005 in the expected direction. Surprisingly, the models containing logForwardTP and MI, higher values in the association metrics correlate with significantly higher RTs. This effect is suggested to be driven by the frequency bias, as it disappears when previous word frequency is controlled for (in addition to noun frequency).

logBackwardTP and logBigramFreq prove to be realistic predictors of reading times; logBigramFreq is the best fit to the data, although logBackwardTP has only a slight disadvantage in terms of model fit (AIC) and proportion of variance accounted for by the fixed effects (marginal R2). The significance of logBigramFreq provides support for the idea of chunk-level activation, as suggested by usage-based approaches. logBackwardTP, on the other hand, suggests concurrent activation of the bigram's component words, though it suggests backwards integration rather than prediction (Blumenthal-Dramé, 2017).

These two 'winner metrics' were additionally compared across two task conditions: the experiment was split into two blocks which only differed in the format of interleaved comprehension questions. In the control block, comprehension questions had a multiple-choice format and in the task block, the same questions (balanced across participants) appeared in a typed free response format. This difference is minimal; free response questions require the reconstruction of knowledge (compared to mere recognition) and have been found to be more difficult (In'nami & Koizumi, 2009).

The multiple-choice condition elicited faster overall reading times and the effects of the two metrics were stronger at the critical word. In comparison, the effect was weaker but longer-lasting across the spillover region in the typed condition, particularly in terms of bigram frequency (see Figure 1). It thus seems possible that during slower reading, low-level information tied to individual words is not abstracted over as quickly, thereby exerting stronger effects on neighboring words (Christiansen & Chater, 2016). We thus argue that insufficient attention to task effects may have partially obscured the cognitive correlates of association scores in similar research.

References

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82. https://doi.org/10.1016/j.jml.2009.09.005

Blumenthal-Dramé, A. (2017). Entrenchment from a psycholinguistic and neurolinguistic perspective. In Hans-Jörg Schmid (Ed.), *Entrenchment and the Psychology of Language Learning: How We Reorganize and Adapt Linguistic Knowledge*. Washington DC: De Gruyter Mouton.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*. https://doi.org/10.1017/S0140525X1500031X

Evert, S. (2009). Corpora and Collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (Vol. 2, pp. 1212–1248). Berlin, New York: Mouton de Gruyter.

Hale, J. (2016). Information-theoretical Complexity Metrics: Information-theoretical complexity metrics. *Language and Linguistics Compass*, *10*(9), 397–412. https://doi.org/10.1111/lnc3.12196

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219–244. https://doi.org/10.1177/0265532208101006

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

The British National Corpus (BNCweb QCP-Edition) (Version 3). (n.d.). Retrieved from http://bncweb.lancs.ac.uk/

Figures

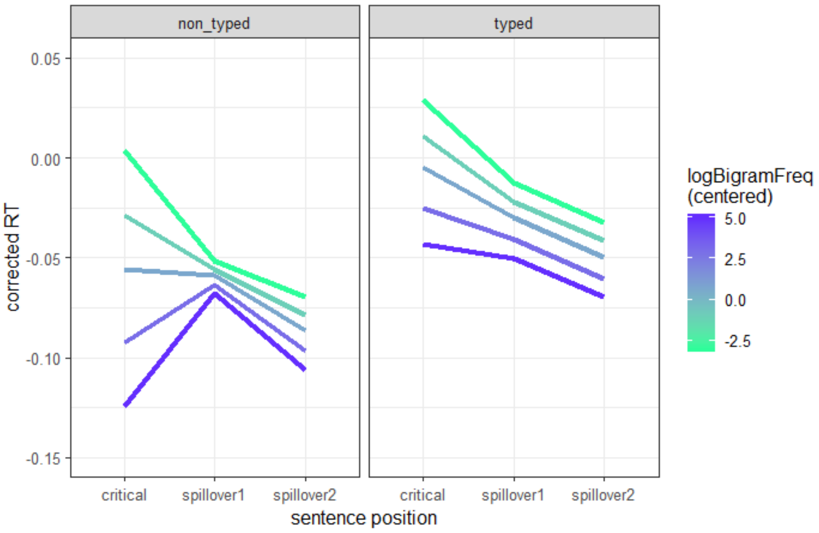

Figure 1: Effects of log-transformed bigram frequency (logBigramFreq) on corrected word-by-word RTs from the critical noun until the second spillover word as a function of required answer format (multiple choice vs. typed free response).